Distribuições de Palavras em sequências aleatórias de letras

Tese de Mestrado orientada pela Doutora Emília Athayde

Maria Luísa Ferreira da Costa Morais

Maio 2004

Conteúdo

Resumo

Neste trabalho pretendemos estudar a distribuição de palavras numa sequência aleatória de letras, que toma valores num alfabeto com N letras. São estudadas a distribuição para a primeira e para a n-ésima ocorrência assim como a distribuição entre ocorrências, tendo por base o caso i.i.d. e o caso Markoviano para a sequência das letras.

A distribuição do número de ocorrências, contadas sem sobreposição, de uma palavra, é aproximada, em determinadas condições, a uma distribuição de Poisson apropriada. Estes resultados utilizam o método de Stein-Chen e são muito importantes uma vez que nos permitem comparar a distribuição do número de ocorrências com uma distribuição de referência.

Os resultados obtidos para a distribuição entre ocorrências sucessivas de palavras são aplicados à análise de sequências de ADN, uma vez que o estudo das distâncias fornece uma maior informação sobre a frequência da palavra mas também sobre a sua distribuição longitudinal ao longo de toda a sequência.

Abstract

In this work we intend to study the distribution of words in a random sequence of letters, taking values in an alphabet with N letters. The distribution of the first occurrence and for the nth occurrence are studied, as well as the distribution between occurrences, under the case i.i.d. and also under the Markovian case for the sequence of the letters.

The distribution of the non-overlapping number of occurences of a word is approximated, under some conditions, to an appropriate Poisson random variable. These results use the Stein-Chen method and are very important because they allow us to compare the distribution of the number of occurrences with a reference random variable.

The results obtained for the distribution between successive occurrences of words are applied to analyse DNA sequences, since the study of the distances give us more information about the frequency of a word but also its longitudinal distribution in the sequence.

Agradecimentos

Ao escrever esta tese perturbei, interrompi, interroguei, mandei mensagens e, no entanto, sempre encontrei paciência, cortesia e acima de tudo disponibilidade. Gostaria de registar aqui as minhas dívidas de gratidão ao João Cortez, ao António e Sandra Pinheiro, à Paula Gomes e à Sandra Varanda e à minha orientadora, Doutora Emília Athayde pela sua competência e disponibilidade na orientação desta tese.

Mas a minha gratidão mais profunda e sentida vai para os meus pais, que ao longo da vida sempre me apoiaram, e para o meu marido que sempre me incentivou nos momentos mais difíceis.

A todos eles o meu muito obrigada.

Capítulo 1

Introdução

Nos campos da Biologia Molecular, da Genética e da Medicina novos resultados estão a surgir a um ritmo impressionante. A primeira fonte de informação consiste na sequenciação do ADN e em consequência na sequenciação de proteínas e macromoléculas. Sequenciar o genoma permitir-nos-á saber muito sobre a evolução dos organismos. Além do objectivo de sequenciar o ADN, é importante descobrir padrões significativos e interpretá-los no processo de formação de proteínas, funções biológicas, desenvolvimento evolutivo, etc. Estes estudos necessitam constantemente de métodos estatísticos. O estudo das distâncias entre ocorrências de "palavras" na sequência de ADN dão-nos muitas informações.

Este problema está intimamente relacionado com o tempo de espera para cadeias de sucessos e com as distribuições de ordem k (distribuições de cadeias de sucessos de comprimento k). O problema original relacionado com o tempo de espera de cadeias de sucessos remontam à era de DeMoivre. A sua aplicação em numerosos campos como a psicologia, metereologia, inferência estatística e controlo de qualidade provocou uma procura contínua e conduziu a diversas extensões do problema original: Feller[2] deu-nos o trabalho clássico sobre cadeias de sucessos; extensões ao caso Markoviano foram propostas; Fu [4] deu-nos o tempo de espera para a n-ésima ocorrência de uma cadeia de sucessos. Robin[8] generalizou o problema, considerando uma sequência de n letras e estudou o tempo de espera para a n-ésima ocorrência, assim como a distribuição entre ocorrências sucessivas,

quer no caso i.i.d, quer tendo por base uma cadeia de Markov. Godbole[6] aproximou o número de ocorrências de uma palavra a uma distribuição de Poisson.

Assim, neste trabalho iremos no capítulo dois, partindo da análise do artigo *"Exact Distribution of word occurrences in a random sequence of letters"* de Robin [8], estudar as distâncias entre ocorrências sucessivas e o tempo de espera para a n-ésima ocorrência. Este estudo é desenvolvido tendo por base o caso i.i.d. e o caso Markoviano [7] quando uma sequência de letras toma valores num alfabeto com mais de duas letras. Deste modo é deduzida uma relação recursiva para as probabilidades nos dois casos, assim como a função geradora de probabilidades e os dois primeiros momentos. O interesse em estudar as distâncias em vez do número de ocorrências advém do facto de se obter um maior número de informações, não só sobre a frequência da palavra, mas também sobre a sua distribuição longitudinal ao longo da sequência.

Como referenciamos anteriormente, o estudo da distribuição entre ocorrências de palavras tem muito interesse, nomeadamente na aplicação à analise das sequências do ADN. Desta forma, no capítulo três vamos aplicar os resultados do capítulo anterior ao estudo de uma sequência do ADN na bactéria *Escherichia coli*. Desta feita foi necessário a criação de um programa informático em *Visual Basic* que permitisse fazer as contagens das ocorrências de palavras. Foi também importante a participação num curso sobre "Análise de sequências de ADN" que decorreu na Universidade do Minho entre 23 e 27 de Julho de 2001, pois permitiu adquirir conhecimentos sobre a importância do ADN, assim como fazer extracções de ADN em bactérias e verificar quais os processos utilizados para fazer a sua sequenciação e respectiva leitura. Ainda no estudo da sequência do ADN da bactéria *Escherichia coli* procurou-se adaptar, da melhor forma possível, os resultados obtidos com os modelos do capítulo dois.

Finalmente no capítulo quatro e tendo por base o artigo de Godbole[6] *"Improved Poisson approximations for word patterns",* iremos provar que o número de ocorrências de uma palavra de comprimento k pode ser aproximado por uma distribuição de Poisson apropriada. Este estudo será desenvolvido quer no caso i.i.d. quer no caso Markoviano e serão utilizados os resultados obtidos por Barbour[1] que têm por base o método Stein-

Chen. Este facto é muito importante uma vez que nos permite comparar a distribuição do número de ocorrências com uma distribuição de referência.

Capítulo 2

Distribuições

2.1 Distribuição Exacta da Ocorrência de Palavras

2.1.1 Introdução

Seja $S = (S_1 S_2 ...)$ uma sequência aleatória de letras que toma valores num alfabeto $\mathcal{A}$ com N letras e seja $W = (w_1 w_2 ... w_k)$ uma palavra de comprimento k.

Diz-se que W ocorre na posição x, se a palavra W acaba na posição x, i.e.,

$\{W \text{ em } x\} = \{S_{x-k+1} ... S_x) = (w_1 ... w_k)\}$

2.1.2 Estrutura repetitiva

A distribuição das ocorrências de W ao longo da sequência depende da própria palavra W. É evidente que o tamanho da palavra assim como a sua estrutura repetitiva tem importância. Esta estrutura repetitiva é dada pelo indicador ε.

Assim $\varepsilon(u) = 1$, se as primeiras u letras da palavra são iguais, e na mesma ordem, às últimas u e $\varepsilon(u) = 0$, caso contrário. Deste modo:

$\varepsilon(u) = 1$ se $\{(w_{k-u+1}...w_k) = (w_1...w_u)\}$ para $1 \leq u \leq k$.

Note-se que se tem sempre $\varepsilon(k) = 1$.

Por exemplo se $W = (gagag)$, então $\varepsilon(u) = 1$ para $u = 1; 3; 5$.

Diz-se que a palavra é não repetitiva se $\varepsilon(u) = 0$ para todos os valores de u excepto para $u = k$. No caso de uma sequência de k letras iguais tem-se $\varepsilon(u) = 1, \text{para } u = 1, ..., k$.

2.1.3 Modelos Estudados

Vamos sucessivamente e ao longo deste trabalho utilizar os seguintes modelos:

Modelo M00

Este modelo trata do caso particular de uma sequência aleatória i.i.d., com distribuição de probabilidade uniforme em $\mathcal{A}$, em que uma letra qualquer tem probabilidade $\frac{1}{N}$ de ocorrer, em qualquer posição x. E assim a palavra W tem probabilidade $\frac{1}{N^k}$, para $x \geq k$.

Modelo M0

Generaliza o modelo anterior, a uma distribuição, que a cada $a \in \mathcal{A}$ associa a probabilidade $\mu(a)$.

Se considerarmos $\tau(u, v) = \prod_{z=u}^{v} \mu(w_z)$, então a probabilidade de W ocorrer, em qualquer posição x, $x \geq k$ pode ser expressa por:

$$\boldsymbol{\mu}(W) = \tau(1, k) = \prod_{z=1}^{k} \mu(w_z)$$

Modelo M1

Este modelo ocorre quando a sequência S é uma cadeia de Markov de primeira ordem com probabilidade de transição π. Em qualquer posição x para quaisquer letras a e b de $\mathcal{A}$, tem-se $\pi(a, b) = P(S_{x+1} = b \,|S_x = a)$.

Assumindo que a sequência está no seu estado estacionário, i.e., em qualquer posição x, a letra a tem probabilidade $\mu(a)$ de ocorrer, sendo μ a distribuição estacionária da cadeia, definimos agora $\tau(u,v)$ como sendo a probabilidade de observarmos $w_u w_{u+1}...w_v$, tendo já ocorrido a letra w_{u-1}, para quaisquer u e v tais que $2 \leq u \leq v \leq k$.

Assim:

$$\tau(u,v) = \prod_{x=u}^{v} \pi(x-1,x)$$

Por convenção $\tau(u+1,u) = 1$.

Em qualquer posição x, com $x \geq k$, a palavra W é observada com probabilidade

$$\boldsymbol{\mu}(W) = \mu(w_1)\tau(2,k)$$

2.2 Distribuição da Primeira Ocorrência

2.2.1 Modelo M00

A probabilidade $p(x)$ de W ocorrer pela primeira vez em x é dada decompondo $P\{W \text{ em } x\}$ em $P\{W \text{ em } x \text{ pela } 1^a\text{vez}\}$ e $P\{W \text{ em } x \text{ não pela primeira vez}\}$ significando esta última que a palavra ocorre em alguma posição $z < x$.

Deste modo, para $x > k$ tem-se que $P\{W \text{ em } x\} = N^{-k}$ uma vez que o processo está no seu estado estacionário.

Se "W está em x não pela primeira vez", significa que pode estar pela primeira vez em z para $z \leq x-k$ com probabilidade $p(z)N^{-k}$ ou então para $x-k < z < x$ com probabilidade $p(z)\varepsilon(k-x+z)N^{z-x}$. Neste último caso, se existir repetição na palavra, i.e, se $\varepsilon(k-x+z) = 1$, as primeiras $k-x+z$ letras são dadas e apenas é necessário observar a sequência $w_{k-x+z+1}...w_k$ que ocorre com probabilidade N^{z-x}.

E assim tem-se, para $x > k$,

$$p(x) = N^{-k} - N^{-k}\sum_{z=k}^{x-k} p(z) - \sum_{z=x-k+1}^{x-1} p(z)\varepsilon(k-x+z)N^{z-x} \qquad (2.1)$$

Note-se que $p(x) = 0$ se $x < k$ e $p(k) = N^{-k}$.

2.2.2 Modelo M1

Teorema 1: No modelo M1 a distribuição da primeira ocorrência é dada por:

$$\begin{aligned} p(x) &= \boldsymbol{\mu}(W) - \sum_{z=k}^{x-k} p(z)\pi^{x-z-k+1}(w_k, w_1)\tau(2,k) - \\ &\quad - \sum_{z=x-k+1}^{x-1} p(z)\varepsilon(k-x+z)\tau(k-x+z+1,k) \end{aligned} \qquad (2.2)$$

com $p(x) = 0$ para $x < k$ e $p(k) = \boldsymbol{\mu}(W)$.

Demonstração: Como no caso anterior $P\{W \text{ em } x\} = \boldsymbol{\mu}(W)$ e W em x não pela primeira vez significa que pode estar em z pela primeira vez para $z \leq x-k$ ou em z pela primeira vez para $x-k < z < x$. A primeira probabilidade será de passar de w_k para w_1 em $x-z-k+1$ passos, o que ocorre com probabilidade $\pi^{x-z-k+1}(w_k, w_1)$ e depois observar $w_2 w_3 ... w_k$ que ocorre com probabilidade $\tau(2,k)$.

No segundo caso, se existir repetição na palavra $[\varepsilon(k-x+z) = 1]$ as primeiras $k-x+z$ letras são dadas e apenas é necessário observar a sequência $w_{k-x+z+1} w_{k-x+z+2} ... w_k$ o que ocorre com probabilidade $\tau(k-x+z+1,k)$.

2.2.3 Função Geradora

Vamos estudar a função geradora de probabilidades (f.g.p.) da variável aleatória X que representa a posição em que W ocorre pela primeira vez ao longo da sequência.

Deste modo temos $\phi_X(t) = E(t^X)$.

Teorema 2: No modelo M00 a função geradora é dada por:

$$\phi_X(t) = \frac{1}{\left[1 + (1-t)\,\delta\left(\frac{1}{t}\right)\right]} \tag{2.3}$$

sendo $\delta(t) = \sum_{u=1}^{k} \varepsilon(u)(Nt)^u$.

Demonstração: De facto pela relação recursiva (2.1) tem-se:

$$\begin{aligned}
\phi_X(t) &= \sum_{x\geq k}\left(N^{-k} - N^{-k}\sum_{z=k}^{x-k} p(z) - \sum_{z=x-k+1}^{x-1} p(z)\varepsilon(k-x+z)N^{z-x}\right)t^x \\
&= N^{-k}\sum_{x\geq k} t^x - N^{-k}\sum_{x\geq k} t^x \sum_{z=k}^{x-k} p(z) - \sum_{x\geq k} t^x \sum_{z=x-k+1}^{x-1} p(z)\varepsilon(k-x+z)N^{z-x} \\
&= N^{-k}\frac{t^k}{1-t} - N^{-k}\sum_{z\geq k} p(z)t^z \sum_{x\geq z+k} t^{x-z} - \\
&\quad -\sum_{z\geq 1} t^z p(z) \sum_{x=z+1}^{z+k-1} \varepsilon(k-x+z)\left(\frac{N}{t}\right)^{z-x} \\
&= N^{-k}\frac{t^k}{1-t} - N^{-k}\phi_X(t)\frac{t^k}{1-t} - \phi_X(t)\sum_{v=1}^{k-1}\varepsilon(k-v)\left(\frac{N}{t}\right)^{-v}
\end{aligned}$$

Multiplicando ambos os membros por $\left(\frac{N}{t}\right)^k (1-t)$ tem-se:

$$\phi_X(t)\left(\frac{N}{t}\right)^k (1-t) = 1 - \phi_X(t) - \phi_X(t)\sum_{v=1}^{k-1}\varepsilon(k-v)(1-t)\left(\frac{N}{t}\right)^{-v+k}$$

$$\Leftrightarrow \quad \phi_X(t)\sum_{v=1}^{k-1}\varepsilon(k-v)(1-t)\left(\frac{N}{t}\right)^{-v+k} + \phi_X(t)\left(\frac{N}{t}\right)^{k}(1-t) = 1-\phi_X(t)$$

$$\Leftrightarrow \quad \phi_X(t)\left(\frac{N}{t}\right)^{k}(1-t)\left[1+\sum_{v=1}^{k-1}\varepsilon(k-v)\left(\frac{N}{t}\right)^{-v}\right] = 1-\phi_X(t)$$

o que é equivalente a ter $\phi_X(t)(1-t)\delta(t^{-1}) = 1-\phi_X(t)$,
uma vez que $\delta(t) = \sum_{u=1}^{k}\varepsilon(u)(Nt)^u$ e $\varepsilon(k) = 1$,
e assim obtém-se o resultado pretendido $\phi_X(t) = \frac{1}{[1+(1-t)\delta(t^{-1})]}$

Teorema 3: No modelo M1 a função geradora é dada por:

$$\phi_X(t) = \frac{1}{\left[\gamma_{k,1}(t) + (1-t)\,\delta\,(t^{-1})\right]} \tag{2.4}$$

sendo $\delta(t) = \sum_{u=1}^{k}\frac{\varepsilon(u)t^u}{\mu(w_1)\tau(2,u)}$ e $\gamma_{k,1}(t) = \frac{1-t}{t\mu(w_1)}\sum_{z\geq 1}\pi^z(w_k,w_1)t^z$

Demonstração: Como no caso anterior, tendo em conta a definição de $\phi_X(t)$ e o facto de $p(x)$ ser dada recursivamente por (2.2) tem-se:

$$\begin{aligned}\phi_X(t) &= \sum_{x\geq k}\boldsymbol{\mu}(W)t^x - \sum_{x\geq k}\sum_{z=k}^{x-k}p(z)\pi^{x-z-k+1}(w_k,w_1)\tau(2,k)t^x - \\ &\quad - \sum_{x\geq k}\sum_{z=x-k+1}^{x-1}p(z)\varepsilon(k-x+z)\tau(k-x+z+1,k)t^x\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\mu}(W)\frac{t^x}{1-t} - \sum_{z\geq k}\sum_{x\geq z+k} p(z)\pi^{x-z-k+1}(w_k,w_1)\tau(2,k)t^x - \\
&\quad - \sum_{z\geq 1}\sum_{x=z+1}^{z+k-1} p(z)\varepsilon(k-x+z)\tau(k-x+z+1,k)t^x \\
&= \boldsymbol{\mu}(W)\frac{t^x}{1-t} - \phi_X(t)\sum_{v\geq k}\pi^{v-k+1}(w_k,w_1)\tau(2,k)t^v - \\
&\quad -\phi_X(t)\sum_{v=1}^{k-1}\varepsilon(k-v)\tau(k-v+1,k)t^v
\end{aligned}
$$

Multiplicando ambos os membros por $\frac{(1-t)}{t^k\mu(W)}$ tem-se:

$$
\begin{aligned}
\phi_X(t)\frac{(1-t)}{t^k\mu(W)} &= 1 - \phi_X(t)\frac{(1-t)}{t^k\mu(W)}\sum_{v\geq k}\pi^{v-k+1}(w_k,w_1)\tau(2,k)t^v - \\
&\quad -\phi_X(t)\frac{(1-t)}{t^k\mu(W)}\sum_{v=1}^{k-1}\varepsilon(k-v)\tau(k-v+1,k)t^v
\end{aligned}
$$

$$
\Leftrightarrow \quad \phi_X(t)\frac{(1-t)}{t^k\boldsymbol{\mu}(W)}\left\{\begin{array}{c} 1+\sum_{v\geq k}\pi^{v-k+1}(w_k,w_1)\tau(2,k)t^v+ \\ +\sum_{v=1}^{k-1}\varepsilon(k-v)\tau(k-v+1,k)t^v \end{array}\right\} = 1
$$

$$
\Leftrightarrow \quad \phi_X(t)\left\{\begin{array}{c} \frac{(1-t)}{t^k\boldsymbol{\mu}(W)} + \frac{(1-t)}{\boldsymbol{\mu}(W)}\sum_{z\geq 1}\pi^z(w_k,w_1)\tau(2,k)t^{z-1}+ \\ +\frac{(1-t)}{t^k\boldsymbol{\mu}(W)}\sum_{v=1}^{k-1}\varepsilon(k-v)\tau(k-v+1,k)t^v \end{array}\right\} = 1
$$

$$\Leftrightarrow \quad \phi_X(t) \left\{ \begin{array}{c} \frac{(1-t)}{t^k \boldsymbol{\mu}(W)} + \frac{(1-t)}{t\boldsymbol{\mu}(W)} \sum_{z\geq 1} \pi^z(w_k, w_1)\tau(2,k)t^z + \\ + \frac{(1-t)}{t^k \boldsymbol{\mu}(W)} \sum_{u=1}^{k-1} \varepsilon(u)\tau(u+1,k)t^{k-u} \end{array} \right\} = 1$$

$$\Leftrightarrow \quad \phi_X(t) \left\{ \begin{array}{c} \frac{(1-t)}{t\boldsymbol{\mu}(W)} \sum_{z\geq 1} \pi^z(w_k, w_1)\tau(2,k)t^z + \\ + \frac{(1-t)}{\boldsymbol{\mu}(W)} \sum_{u=1}^{k} \varepsilon(u)\tau(u+1,k)t^{-u} \end{array} \right\} = 1$$

uma vez que $\frac{\tau(2,k)}{\tau(2,u)} = \tau(u+1,k)$ e $\boldsymbol{\mu}(W) = \mu(w_1)\tau(2,k)$ tem-se que:

$$\phi_X(t) \left[\gamma_{k,1}(t) + (1-t) \sum_{u=1}^{k} \frac{\varepsilon(u)t^{-u}}{\mu(w_1)\tau(2,u)} \right] = 1$$

Como $\phi_X(t)$ é uma função geradora de probabilidades e consequentemente $\phi_X(1) = 1$ tem-se que $\gamma_{k,1}(1) = 1$.

2.2.4 Momentos

O valor médio e a variância são dados, utilizando as propriedades de $\phi_X(t)$ por:

$E(X) = \phi'_X(1)$ e $V(X) = \phi''_X(1) + \phi'_X(1)[1 - \phi'_X(1)]$

Modelo M00

Neste modelo temos que:

$$E(X) = \delta(1) \text{ e } V(X) = \delta^2(1) - 2\delta^{'}(1) + \delta(1)$$

Este resultado advém do facto de:

$$\begin{aligned} \phi'_X(t) &= \frac{\delta(t^{-1}) + (t^{-2} - t^{-1})\,\delta'(t^{-1})}{[1 + (1-t)\delta(t^{-1})]^2} \\ \phi''_X(t) &= \frac{[-2t^{-3}\delta'(t^{-1}) - (t^{-4} - t^{-3})\delta''(t^{-1})][1 + (1-t)\delta(t^{-1})]}{[1 + (1-t)\delta(t^{-1})]^3} + \\ &\quad + \frac{2[\delta(t^{-1}) + (t^{-2} - t^{-1})\,\delta'(t^{-1})]^2}{[1 + (1-t)\delta(t^{-1})]^3} \end{aligned}$$

Exemplo: No lançamento de uma moeda equilibrada ($N = 2$) o valor médio da primeira ocorrência da palavra $W = (01)$ é 4 assim como a variância, mas para a palavra $W = (11)$ o valor médio é 6 e a variância 22.

Como $\delta(t) = (Nt)^k$ para palavras não repetitivas, o valor médio é sempre maior para palavras repetitivas do que para palavras não repetitivas com o mesmo comprimento. O mesmo se passa para a variância.

Modelo M1

Para determinarmos os momentos neste modelo, é necessário a primeira e a segunda derivada de $\gamma_{k,1}(t)$. Isto pode ser feito utilizando uma outra forma desta função dada pelo seguinte lema:

Lema 1: Seja Π a matriz de probabilidades de transição da cadeia de Markov de tal modo que $\Pi_{a,b} = \pi(a, b)$.

Suponhamos que Π admite uma decomposição $\Pi = A\Lambda B$ em que Λ é uma matriz diagonal com $\lambda_1 = 1$ e $1 > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_N|$ e $AB = BA = I$[1].

Partindo desta hipótese tem-se que

$$\gamma_{k,1}(t) = 1 + \frac{1-t}{\mu(w_1)} \sum_{i=2}^{N} \frac{\lambda_i}{1 - t\lambda_i} a_{w_k,i} b_{i,w_1} \tag{2.5}$$

onde $a_{i,j}$ e $b_{i,j}$ são os elementos genéricos de A e B.

Demonstração: Tem-se que $\gamma_{k,1}(t) = \Gamma_{w_k,w_1}(t)$ sendo $\Gamma(t)$ a matriz definida para todo o $t > 0$ por

$$\Gamma(t) = \frac{1-t}{t\mu(w_1)} \left[(I - t\Pi)^{-1} - I \right] = \frac{1-t}{t\mu(w_1)} \sum_{z \geq 1} (t\Pi)^z$$

sendo I a matriz identidade de dimensão N.

De facto $(I - t\Pi) \left(I + \sum_{z \geq 1} (t\Pi)^z \right) = \left(I + \sum_{z \geq 1} (t\Pi)^z \right) (I - t\Pi) = I$

[1] A matriz Λ é a matriz diagonal em que os seus elementos são os valores próprios de Π. A matriz A é a matriz dos vectores próprios (em coluna) e B é a inversa de A

Usando a hipótese da matriz Π admitir a decomposição acima descrita tem-se que:[2]

$$\begin{aligned}\Gamma(t) &= \frac{1-t}{t\mu(w_1)} A\left[(I-t\Lambda)^{-1}-I\right]B = \frac{1-t}{t\mu(w_1)}\sum_{i=1}^{N}\left(\frac{1}{1-t\lambda_i}-1\right)a_i b_i \\ &= \frac{1-t}{\mu(w_1)}\sum_{i=1}^{N}\frac{\lambda_i}{1-t\lambda_i}a_i b_i\end{aligned}$$

denotando-se a_i a i-ésima coluna de A e b_i a i-ésima linha de B de tal modo que $(a_i b_i)_{u,v} = a_{u,i}b_{j,v}$, sendo $a_{u,i}$ o elemento na coluna i e na linha correspondente à letra w_u da matriz A.

Note-se que $a_1 = [1\,1\,...\,1]^T$ e $b_1 = \mu$ (sendo μ a probabilidade estacionária).

Usando o facto de $\lambda_1 = 1$ tem-se que:

$$\Gamma(t) = \frac{a_1 b_1}{\mu(w_1)} + \frac{1-t}{\mu(w_1)}\sum_{i=2}^{N}\frac{\lambda_i}{1-t\lambda_i}a_i b_i$$

Desta forma o lema está demonstrado e podemos então determinar a primeira e a segunda derivada de $\gamma_{k,1}$.

$$\begin{aligned}\gamma_{k,1}^{'}(t) &= \frac{-1}{\mu(w_1)} + \sum_{i=2}^{N}\frac{\lambda_i}{1-t\lambda_i}a_i b_i + \frac{1-t}{\mu(w_1)}\sum_{i=2}^{N}\frac{\lambda_i^2}{(1-t\lambda_i)^2}a_i b_i \\ \gamma_{k,1}^{''}(t) &= \frac{-2}{\mu(w_1)}\sum_{i=2}^{N}\frac{\lambda_i^2}{(1-t\lambda_i)^2}a_i b_i + 2\frac{1-t}{\mu(w_1)}\sum_{i=2}^{N}\frac{\lambda_i^3}{(1-t\lambda_i)^3}a_i b_i\end{aligned}$$

[2]De facto $A[(I-t\Lambda)^{-1}-I]B = B^{-1}(I-t\Lambda)^{-1}A^{-1} - I = [A(I-t\Lambda)B]^{-1} - I = (I-t\Pi)^{-1} - I$,

Para $t = 1$, tem-se:

$$\begin{aligned} \gamma_{k,1}^{'}(1) &= \frac{-1}{\mu(w_1)} + \sum_{i=2}^{N} \frac{\lambda_i}{1-\lambda_i} a_i b_i \\ \gamma_{k,1}^{''}(t) &= \frac{-2}{\mu(w_1)} \sum_{i=2}^{N} \frac{\lambda_i^2}{(1-\lambda_i)^2} a_i b_i \end{aligned}$$

Encontra-se ainda uma relação muito simples para a m-ésima derivada de $\gamma_{k,1}(t)$ no ponto $t = 1$.

$$\gamma_{k,1}^{(m)}(1) = \frac{-m!}{\mu(w_1)} \sum_{i=2}^{N} \frac{\lambda_i^m}{(1-\lambda_i)^m} a_{w_{k,}i} b_{i,w_1}$$

Podemos então encontrar os momentos para X no modelo M1

$$\begin{aligned} E(X) &= \delta(1) - \gamma_{k,1}^{'}(t) \\ V(X) &= \left[\delta(1) - \gamma_{k,1}^{'}(1)\right]^2 - 2\delta'(1) - \gamma_{k,1}^{''}(t) + \left[\delta(1) - \gamma_{k,1}^{'}(1)\right] \end{aligned}$$

Exemplo: Sendo $\mathcal{A} = \{0, 1\}$ e $\Pi = \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ tem-se:

O valor médio da primeira ocorrência da palavra $W = (01)$ é $\frac{11}{2}$ e a variância $\frac{45}{4}$ enquanto que se $W = (11)$ o valor médio é o mesmo, mas a variância passa a ser $\frac{81}{4}$. Em palavras repetitivas a variância é maior do que em palavras não repetitivas.

2.3 Distribuição da distância entre duas ocorrências sucessivas

2.3.1 Modelo M00

Iremos agora estudar a distância Y entre duas ocorrências sucessivas de W, uma na posição x_0 e a outra na posição $x_0 + Y$

Teorema 4: No modelo M00 a probabilidade $q(y) = P(Y = y)$ é dada da seguinte forma:

se $0 < y < k$, então

$$q(y) = N^{-y}\varepsilon(k-y) - \sum_{z=1}^{y-1} q(z)\varepsilon(k-y+z)N^{z-y} \tag{2.6}$$

se $y \geq k$, tem-se

$$q(y) = N^{-k} - N^{-k}\sum_{z=1}^{y-k} q(z) - \sum_{z=y-k+1}^{y-1} q(z)\varepsilon(k-y+z)N^{z-y} \tag{2.7}$$

Demonstração: Para $y \geq k$, a fórmula recursiva é a mesma que (2.1). Para $0 < y < k$, usa-se a mesma decomposição mas, $P(W$ em $x_0 + y \,|W$ em $x_0)$ é $N^{-y}\varepsilon(k-y)$ em vez de N^{-k}, uma vez que a ocorrência repete-se necessariamente na anterior. O outro termo é análogo ao caso $y \geq k$, mas a primeira soma é zero.

2.3.2 Modelo M1

Teorema 5: No modelo M1, a distribuição da distância Y entre duas sucessivas ocorrências da palavra tem a relação recursiva dada por:

se $0 < y < k$, então

$$q(y) = \varepsilon(k-y)\tau(k-y+1,k) - \sum_{z=1}^{y-1} q(z)\varepsilon(k-y+z)\tau(k-y+z+1,k)$$

se $y \geq k$, tem-se:

$$\begin{aligned} q(y) &= \pi^{y-k+1}(w_k, w_1)\tau(2,k) - \sum_{z=1}^{y-1} q(z)\pi^{y-z-k+1}(w_k, w_1)\tau(2,k) - \\ &\quad - \sum_{z=y-k+1}^{y-1} q(z)\varepsilon(k-y+z)\tau(k-y+z+1,k) \end{aligned} \tag{2.8}$$

Demonstração: Para $y \geq k$, a fórmula recursiva é a mesma que (2.2), apenas em vez de $\boldsymbol{\mu}(W)$ tem-se $\pi^{y-k+1}(w_k, w_1)\tau(2,k)$, uma vez que a palavra tem de transitar de w_1 para w_k em $y-k+1$ passos. Para $0 < y < k$, usa-se a mesma decomposição, mas $P(W$ em $x_0 + y \,|W$ em $x_0)$ é $\varepsilon(k-y)$ $\tau(k-y+1,k)$ em vez de $\boldsymbol{\mu}(W)$, uma vez que a ocorrência repete-se necessariamente na anterior. O outro termo é análogo ao caso $y \geq k$, mas a primeira soma é zero.

Note-se que, mesmo quando $\varepsilon(k-y) = 1$, $q(y)$ pode ser nula pelo que algumas distâncias correspondentes a sobreposições são impossíveis.

2.3.3 Função Geradora

A função geradora de probabilidades da distância Y é dada, determinando a série $\sum_{y\geq 1} q(y)t^y$.

Deste modo no modelo M00 tem-se que:

$$\phi_Y(t) = 1 - \frac{(1-t)(N/t)^k}{1 + (1-t)\delta(t^{-1})} \tag{2.9}$$

sendo $\delta(t)$ definido como anteriormente.
No modelo M1 tem-se

$$\phi_Y(t) = 1 - \frac{(1-t)\boldsymbol{\mu}(W)^{-1}t^{-k}}{\gamma_{k,1}(t) + (1-t)\delta(t^{-1})} \tag{2.10}$$

sendo $\delta(t)$ e $\gamma_{k,1}(t)$ de igual modo definidos como anteriormente.

2.3.4 Momentos

Como anteriormente podemos determinar os dois primeiros momentos.
Desta forma por (2.9) determinamos

$$E_{\mathrm{M00}}(Y) = N^k \text{ e } V_{\mathrm{M00}}(Y) = N^k\left[2\delta(1) - (2k-1)\right] - N^{2k}$$

Verificamos que o valor médio apenas depende do comprimento da palavra, mas tal não é verdade para a variância.

Por (2.10) e utilizando (2.5) tem-se:

$$E_{\mathrm{M1}}(Y) = \frac{1}{\mu(W)}$$

$$V_{\mathrm{M1}} = \frac{1}{\mu(W)}\left\{2\left[\delta(1) - \gamma'_{k,1}(1)\right] - (2k-1)\right\} - \left(\frac{1}{\mu(W)}\right)^2$$

2.4 Distribuição da n-ésima ocorrência

2.4.1 Modelo M00

Teorema 8: No modelo M00 a distribuição da posição X_n da n-ésima ocorrência da palavra W é, para $x \geq k$

$$p(x,n) = N^{-k} - N^{-k}\sum_{z=1}^{x-k} p(z,n) - \sum_{z=x-k+1}^{x-1} p(z,n)\varepsilon(k-x+z)N^{z-x} - \\ -\sum_{m=1}^{n-1} p(x,m) \tag{2.11}$$

sendo $p(x,n) = 0$ para $x < k+n-1$ e $p(k,1) = N^{-k}$
(por convenção $p(.,0) = 0$).

Demonstração: O princípio é o mesmo utilizado em (2.1), (2.6) e (2.7), onde N^{-k} é a probabilidade de W ocorrer na posição x e $p(x,n)$ é a probabilidade de ocorrer em x pela n-ésima vez. As duas primeiras somas são os casos em que essa ocorrência já aconteceu anteriormente em z ($z \leq x-k$ ou $x-k < z < x$) e o último termo é a probabilidade da m-ésima ocorrência, com $m < n$.

2.4.2 Modelo M1

Teorema 9: A distribuição da n-ésima ocorrência de W na posição X_n é dada, para $x \geq k$ por:

$$p(x,n) = \mu(W) - \sum_{z=1}^{x-k} p(z,n)\pi^{x-z-k+1}(w_k, w_1)\tau(2,k) - \\ - \sum_{z=x-k+1}^{x-1} p(z,n)\varepsilon(k-x+z)\tau(k-x+z+1,k) - \\ -\sum_{m=1}^{n-1} p(x,m) \tag{2.12}$$

sendo $p(x,n) = 0$ para $x < k+n-1$ e $p(x,1) = p(x)$
(por convenção $p(x,.) = 0$ para $x \leq 0$)

Demonstração: O princípio é o mesmo utilizado em (2.11)

2.4.3 Função Geradora

Teorema 10: A função geradora de X_n é:

$$\phi_{X_n}(t) = \phi_X(t) \times [\phi_Y(t)]^{n-1} \tag{2.13}$$

Demonstração: O resultado advém do facto de $X_n = X + \sum_{m=1}^{n-1} Y_m$ e, de tanto no modelo M00 como no modelo M1, $X, Y_1, ..., Y_{n-1}$ serem mutuamente independentes.

Desta forma

$$\begin{aligned} \phi_{X_n}(t) &= E(t^{X+Y_1+...+Y_{n-1}}) = E(t^X) \times E(t^{Y_1}) \times ... \times E(t^{Y_{n-1}}) \\ &= \phi_X(t) \times \phi_Y(t) \times ... \times \phi_Y(t) \\ &= \phi_X(t) \times [\phi_Y(t)]^{n-1} \end{aligned}$$

Segue também que:

$$\begin{aligned} E(X_n) &= \phi'_{X_n}(1) = E(X) + (n-1)E(Y) \\ V(X_n) &= \phi''_{X_n}(1) + \phi'_{X_n}(1)\left[1 - \phi'_{X_n}(1)\right] = V(X) + (n-1)V(Y) \end{aligned}$$

2.5 Distribuição da distância entre a n-ésima e a (n+r)-ésima ocorrência

Define-se a distância entre a n-ésima ocorrência e a (n+r)-ésima ocorrência da palavra W como $Y_r = \sum_{j=1}^{r} Y'_j$ com Y'_j a distância entre a (n+j-1)-ésima e a (n+j) -ésima ocorrência, com a distribuição de Y, definida como anteriormente.

2.5.1 Modelo M00

Neste modelo a probabilidade $q(y,r) = \Pr(Y_r = y)$ é dada, para $y \geq k$ pela fórmula recursiva

$$\begin{aligned} q(y,r) &= N^{-k} - N^{-k}\sum_{z=1}^{y-k} q(z,r) - \sum_{z=y-k+1}^{y-1} q(z,r)\varepsilon(k-y+z)N^{z-y} - \\ &\quad -\sum_{m=1}^{r-1} q(y,m) \end{aligned} \tag{2.14}$$

sendo $q(y,r) = 0$ para $y < r$ e $q(y,1) = q(y)$
(por convenção $p(y,.) = 0$ para $y \leq 0$)

2.5.2 Modelo M1

No modelo M1 a relação é dada para $y \geq k$ por:

$$\begin{aligned} q(y,r) &= \pi^{y-k+1}(w_k, w_1)\tau(2,k) - \sum_{z=1}^{y-k} q(z,r)\pi^{y-z-k+1}(w_k, w_1)\tau(2,k) - \\ &\quad - \sum_{z=y-k+1}^{y-1} q(z,r)\varepsilon(k-y+z)\tau(k-y+z+1,k) - \\ &\quad -\sum_{m=1}^{r-1} q(y,m) \end{aligned} \tag{2.15}$$

sendo $q(y,r) = 0$ para $y < r$ e $q(y,1) = q(y)$
(por convenção $q(y,.) = 0$ para $y \leq 0$).

2.5.3 Função Geradora

Nos dois modelos e uma vez que as distâncias $Y_1^{'}, ..., Y_r^{'}$ são mutuamente independentes e identicamente distribuídas, temos que:

$$\phi_{Y_r}(t) = [\phi_Y(t)]^r$$

É obvio também que:

$$E(Y_r) = rE(Y) \text{ e } V(Y_r) = rV(Y)$$

2.6 Distância entre duas palavras

Os resultados anteriores poderão ser adaptados no caso da distância Z entre duas palavras V e W quaisquer, nesta ordem e assumindo que V não está contida em W e vice-versa. Considere-se $W^*(w_1...w_{k^*})$ a maior "sub-palavra" comum a V e W no sentido em que V termina com W^* e W começa com W^*(não existindo outra maior nas mesmas condições).

No modelo M00 dados (2.6) e (2.7) apenas temos de substituir o primeiro $\varepsilon(k-y)$ por $\varepsilon^*(k-y)$, sendo ε^* o indicador de repetição de W sobre W^* com u letras em comum. O outro ε é o indicador de repetição da palavra W.

No modelo M1 sendo

$$\delta^*(t) = \sum_{u=1}^{k^*} \frac{\varepsilon^*(u)t^u}{\mu(w_1)\tau(2,u)} \text{ e } \delta(t) = \sum_{u=1}^{k} \frac{\varepsilon(u)t^u}{\mu(w_1)\tau(2,u)}$$

tem-se:

$$\phi_Z(t) = \frac{\gamma_{k^*,1}(t) + (1-t)\delta^*(t^{-1})}{\gamma_{k,1}(t) + (1-t)\delta(t^{-1})}$$

Desta forma

$$E_{\mathrm{M1}}(Z) = \delta(1) - \delta^*(1)$$

$$\begin{aligned} V_{\mathrm{M1}}(Z) &= [\mu(W)]^{-1}\left\{2\left[\delta(1)-\gamma'_{k,1}(t)\right]-(2k-1)-[\mu(W)]^{-1}\right\}+ \\ &\quad +[\mu(W^*)]^{-1}\left\{2\left[\delta^*(1)-\gamma'_{k^*,1}(t)\right]-(2k^*-1)-[\mu(W^*)]^{-1}\right\} \end{aligned}$$

Uma vez que Z é a posição da primeira ocorrência de W depois de V, podemos utilizar os resultados anteriores para definir, de um modo análogo, a distribuição para a n-ésima ocorrência.

Capítulo 3

Aplicação ao Estudo do ADN

Vamos agora aplicar estes modelos, em especial o modelo M1 ao estudo de sequências de ADN. Desde muito cedo se tem trabalhado na descoberta e interpretação de sequências de ADN para entender o mecanismo molecular e o processo evolutivo. Existem vários programas mundiais que estudam o genoma de várias espécies, em especial o genoma humano. Iremos aplicar este modelo a uma bactéria - *Escherichia coli* - que é muito estudada pelos cientistas, uma vez que toda a sua sequência de ADN é conhecida desde 1997. Mas antes de mais iremos referir alguns conhecimentos básicos sobre o ADN e também dar algumas informações sobre a bactéria estudada.

3.1 Noções básicas sobre o ADN

Porque diferem os seres vivos uns dos outros?

Os biólogos identificaram já cerca de dois milhões de espécies, mas a total diversidade de vida está estimada, em cerca de trinta a quarenta milhões. Uma explicação para esta diversidade reside no ADN, ácido desoxirribonucleico. Esta biomolécula é o suporte molecular de informação biológica que define as caracteríticas de cada organismo. O ADN pertence ao grupo dos ácidos nucleicos, as biomoléculas mais importantes do

controlo celular, pois contêm a informação genética. Existem dois tipos de ácidos nucleicos, o ácido desoxirribonucleico (ADN) e o ácido ribonucleico (ARN).

Os ácidos nucleicos são polímeros em que a unidade básica que os constitui são chamados nucleótidos. Cada nucleótido é constituído por três componentes diferentes: um grupo fosfato, uma pentose e uma base azotada, sendo designado pela base que entra na sua constituição. Existem cinco bases azotadas: Adenina(A); Timina(T); Guanina(G); Citosina(C) e Uracilo(U) e por consequente cinco categorias de nucleótidos. A Timina só está presente no ADN e o Uracilo apenas no ARN. Em cada espécie, o número de adeninas é próximo do número de tininas e o número de guaninas é próximo do número de citosinas. Assim $\frac{n_A+n_G}{n_C+n_T} \simeq 1$ (sendo n_A o número de adeninas, etc.)

A ordem em que aparecem as bases na molécula de ADN, constitui as instruções do programa genético dos organismos. Conhecer a sequência das bases, isto é, sequenciar o ADN equivale a decifrar a informação genética de um organismo - genoma.

Figura 3.1: Estrutura do ADN

Em 1953 James Watson e Francis Crick apresentam uma proposta de modelo em dupla hélice para a estrutura de uma molécula de ADN. Cada molécula de ADN é constituída por duas cadeias enroladas helicoidalmente à volta de um mesmo eixo. As bandas laterais são cadeias formadas por fosfato e desoxirribose (pentose) e os "degraus" centrais são pares de bases ligados entre si por ligações de hidrogénio. A adenina liga-se por duas ligações de hidrogénio à timina e a guanina liga-se à citosina por três ligações de hidrogénio.

A estrutura de dupla hélice do ADN implica que a sequência das bases de uma das cadeias delimita automaticamente a ordem na outra - são cadeias complementares. Uma vez conhecida a sequência das bases de uma cadeia deduz-se automaticamente a sequência das bases na sua complementar. Assim a sequência GATC torna-se CTAG na complementar.

A informação genética encontra-se codificada na sequência de nucleótidos que constituem o ADN. Como o material genético tem a capacidade de ser copiado, as informações são transmitidas de uma geração à geração seguinte, através de um processo de replicação que assegura a conservação e transmissão do património genético próprio de cada espécie. Neste processo, a dupla hélice separa-se e cada cadeia serve de molde para a síntese de uma nova cadeia. O resultado final são duas moléculas idênticas à original.

A molécula de ADN apresenta uma organização e funcionamento universal em todos os seres vivos. Nos procariontes, o ADN encontra-se disperso no citoplasma apresentando-se como uma molécula única circular enquanto que nos seres eucariontes encontra-se essencialmente no núcleo e faz parte de estruturas designadas por cromossomas. Por exemplo a bactéria *Escherichia coli* tem apenas um cromossoma circular enquanto que os humanos têm vinte e três pares de cromossomas lineares. No entanto, em qualquer um dos casos, o ADN é o suporte universal da informação genética, controlando toda a informação hereditária que passa de geração em geração.

A expressão hereditária verifica-se, por exemplo, ao nível da síntese proteica. A unidade básica de estrutura das proteínas são monómeros designados por aminoácidos, ordenados numa sequência linear particular. Existem cerca de vinte aminoácidos comuns a todos os organismos. A ordenação dos aminoácidos numa proteína confere-lhes características bi-

ológicas muito específicas reflectindo-se essas características nas especificidades inerentes a cada ser vivo. A alteração de um aminoácido numa proteína pode conduzir a uma modificação no comportamento e função biológica dessa proteína.

A informação para a ordenação dos aminoácidos está codificada no ADN sob a forma de um código que reside na sequência das suas bases. Os investigadores apontam para um código de três nucleótidos consecutivos de ADN, que representa a mais pequena unidade da mensagem genética - aminoácido. Esta informação de ADN é transmitida para uma fracção de ARN. A cada tripleto (conjunto de três bases) do ADN corresponde, no ARN, um tripleto de bases complementares. Assim, por exemplo, a uma sequência AAA do ADN, corresponde no ARN ao tripleto UUU (o ARN utiliza Uracilo em vez de Timina), tripleto este designado de codão e que codifica o aminoácido fenilalanina. Desta forma podemos afirmar que a sequência AAA no ADN transporta a informação para a codificação desse aminoácido.

Na tabela seguinte apresentamos as sequências de ADN e o respectivo aminoácido que essa sequência codifica quando é transmitida ao ARN. Sequências diferentes de ADN transportam a informação para a sintetização do mesmo aminoácido. No máximo um aminoácido pode ser codificado por seis sequências diferentes.

Fenilalanina	AAA	AAG				
Leucina	AAT	AAC	GAA	GAG	GAT	GAC
Isoleucina	TAA	TAG	TAT			
Valina	CAA	CAG	CAT	CAC		
Serina	AGA	AGG	AGT	AGC	TCA	TCG
Prolina	GGA	GGG	GGT	GGC		
Treonina	TGA	TGG	TGT	TGC		
Alanina	CGA	CGG	CGT	CGC		
Tirosina	ATA	ATG				
Arginina	GCA	GCG	GCT	GCC		
Glicina	CCA	CCG	CCT	CCC		

Asparagina	TTA	TTG
Lisina	TTT	TTC
Cisteína	ACC	ACG
Glutamina	GTT	GTC
Histidina	GTA	GTG
Ácido aspartámico	CTA	CTG
Ácido glutâmico	CTT	CTC
Triptofano	ACC	
Metionina	TAC	
Terminação	ATT	ATC

A grande diversidade de moléculas de ADN confere grande diversidade à vida, pois cada organismo contém o seu ADN, que o torna único.

3.2 Escherichia coli

A *Escherichia coli* é uma bactéria que está presente nas plantas e na flora intestinal de todos os animais. Além disso é a bactéria mais encontrada nos laboratórios e a mais estudada. Pertence à famíla *Enterobacteraceae* e o seu nome deve-se ao físico alemão Theodor Escherich, que em 1885 a isolou e caracterizou pela primeira vez. Fazendo parte da flora intestinal dos humanos, a *E. coli* tem um papel crucial na digestão. No entanto, espécies patogénicas da bactéria podem causar sérios casos de diarreias, meningites, pneumonias, etc.

Figura 3.2: *Escherichia coli*

3.3 Exemplo

Iremos estudar a sequência "ecomori"[1] que é uma parte do genoma da Escherichia coli e cujo comprimento é 111416.

Utilizaremos o modelo M1 nesta sequência e com a matriz de probabilidades de transição estimada[2]

$$\Pi = \begin{array}{c|cccc} & \text{a} & \text{g} & \text{c} & \text{t} \\ \hline \text{a} & 0.2927 & 0.2182 & 0.2219 & 0.2672 \\ \text{g} & 0.2208 & 0.2428 & 0.3223 & 0.2141 \\ \text{c} & 0.2558 & 0.3201 & 0.2228 & 0.2014 \\ \text{t} & 0.1695 & 0.3157 & 0.2309 & 0.2838 \end{array}$$

Se considerarmos a palavra $W = (gatc)$, este modelo dá-nos:

$\boldsymbol{\mu}(W) = 0.003734$, $E(Y) = 267.9$ e $V(Y) = (265.4)^2$ [3]. O número esperado de ocorrências de W seria então de $111416/E(Y) = 415.9$, enquanto que na sequência existem 495 ocorrências de W. Isto far-nos-á pensar que o modelo não está muito bem adequado ao estudo das sequências do ADN.

Se considerarmos uma alteração ao modelo, tornando-o mais refinado, em que as probabilidades de transição têm em conta os dois instantes anteriores, teremos então uma matriz de probabilidades de transição de dimensão 16. Desta forma consideremos a matriz de transição estimada:

---

[1] Esta sequência encontra-se disponível no site www.colibri.com

[2] As contagens das letras nesta sequência foram feitas utilizando um programa em *Visual Basic* criado para o efeito.

[3] Ver em Apêndice a decomposição da matriz Π em $A\Lambda B$.

$\Pi =$

aa	ag	ac	at	ga	gg	gc	gt	ca	cg	cc	ct	ta	tg	tc	tt
.32	.21	.24	.23	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	.24	.22	.34	.2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.2	.3	.29	.2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	.0	.17	.28	.29	.26
.32	.15	.21	.32	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	.19	.18	.36	.28	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.24	.33	.22	.21	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	.19	.29	.21	.31
.24	.34	.19	.23	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	.21	.26	.31	.22	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.3	.36	.17	.18	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	.11	.48	.17	.24
.3	.14	.26	.31	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	.24	.3	.29	.16	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	.28	.29	.22	.21	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	.19	.25	.24	.31

A tabela seguinte mostra os resultados dos dois modelos para os vinte aminoácidos diferentes. Apenas são apresentados os resultados para uma das várias sequências que codificam determinado aminoácido, no entanto para todas as outras sequências os resultados são melhores com o modelo refinado.

Aminoácido	Sequência	n^{o}ocorrências	n^{o}esperado modelo M1	n^{o}esperado modelo refinado
Fenilalanina	AAA	2449	2233	2430
Leucina	GAC	2151	1971	2133
Isoleucina	TAA	1374	1326	1357
Metionina	TAC	1165	1006	1174
Valina	CAT	1664	1918	1653
Serina	TCG	1810	1976	1796
Prolina	GGT	2046	1587	2058
Treonina	TGA	2093	1862	2099
Alanina	CGA	1897	1982	1906
Tirosina	ATG	1961	2200	1943
Histidina	GTG	1891	2062	1902
Glutamina	GTC	1361	1508	1369
Asparagina	TTA	1464	1286	1460
Ácido aspartâmico	CTA	597	958	600
Ácido glutâmico	CTC	983	1304	988
Lisina	TTT	2379	2155	2374
Cisteína	ACG	1741	1851	1742
Arginina	GCT	2077	1981	2068
Glicina	CCC	1047	1394	1044
Triptofano	ACC	1700	1290	1701

Os resultados são muito melhores para sequências de três letras - que têm informação para codificar os aminoácidos. Pelo que neste caso, a alteração ao modelo conduz a resultados significativos.

Capítulo 4

Aproximações

Muito trabalho tem sido feito para encontrar palavras com frequências inesperadas utilizando para tal aproximações à distribuição de Poisson. O estudo das distâncias entre ocorrências de palavras está relacionado com distribuições de cadeias de sucessos. O clássico trabalho no estudo de cadeias de sucessos deve-se a Feller [2]. Fu e Koutras [4] apresentaram aproximações para as estatísticas mais utilizadas no estudo das cadeias de sucessos, utilizando uma técnica que consiste em "embeber" as variáveis em cadeias de Markov. Uma das estatísticas estudadas foi $N = N(n, k)$ que representa o número de cadeias de k sucessos contados sem sobreposição, isto é, no sentido de Feller. Apresentaram ainda a distribuição do tempo de espera para a n-ésima ocorrência de uma cadeia de k sucessos.

Iremos então neste capítulo estudar a distância de variação total entre a variável N e uma distribuição de Poisson apropriada, quer no caso i.i.d., quer no caso Markoviano. Para tal iremos utilizar os resultados obtidos por Barbour[1] que têm por base o método de Stein-Chen.

Consideremos então uma sequência $X_1, X_2, ..., X_n$ de variáveis de Bernoulli independentes com probabilidade p e seja $N = N(n, k)$ o número de ocorrências não sobrepostas de uma sequência de sucessos de comprimento k. Alternativamente $X_1, X_2, ..., X_n$ pode ser uma sequência de variáveis i.i.d. com valores num alfabeto $\mathcal{A}$ com ξ letras e sendo $N = N(n, k)$ o número de ocorrências contadas sem sobreposição de um palavra fixa de

comprimento k. Este esquema também pode ser utilizado tendo por base uma cadeia de Markov de dois ou ξ estados. Em qualquer caso é razoável esperar que a distribuição de N possa ser aproximada por uma distribuição de Poisson, desde que uma única ocorrência da palavra seja considerado um acontecimento raro.

4.1 Caso i.i.d.

Iremos então considerar o caso em que temos uma sequência aleatória i.i.d. que toma valores no alfabeto $\mathcal{A} = \{I, S\}$ e consideremos as seguintes variáveis:

A : número de ocorrências da palavra $IS...S$ (k $S's$)

B : número de cadeias de sucessos de comprimento pelo menos k.

$N = N(n, k)$: número de ocorrências de k sucessos consecutivos contados sem sobreposição.

$d(.,.)$: distância de variação total, isto é:

$$d(W, Y)) = \sup_C |P(W \in C) - P(Y \in C)|$$

$Po(\lambda)$: variável com distribuição de Poisson com parâmetro λ.

Temos então que:

$$\begin{aligned} d(N, B) &\leq P(N \neq B) = \\ &= P(\text{existir uma cadeia de pelo menos } 2k \text{ sucessos}) \\ &\leq np^{2k} \end{aligned} \tag{4.1}$$

$$\begin{aligned} d(A, B) &\leq P(A \neq B) = \\ &= P(\text{os primeiros } k \text{ serem todos sucessos}) = p^k \end{aligned} \tag{4.2}$$

Utilizando o Teorema II.C de Barbour[1]:

$$d(A, Po(E(A)) \leq (2k + 1)qp^k \tag{4.3}$$

tem-se então:

$$\begin{aligned} d(N, Po(E(A)) &\leq d(N, B) + d(B, A) + d(A, Po(E(A)) \\ &\leq (2k+2)p^k + np^{2k} \end{aligned} \tag{4.4}$$

Esta aproximação é de ordem $O(kp^k)$, quando $np^k \underset{n\to\infty}{\longrightarrow} \lambda$.

4.2 Caso Markoviano

Iremos agora tratar da generalização do problema, assumindo que as letras são geradas percorrendo uma cadeia de Markov com probabilidade estacionária μ e matriz de probabilidades de transição Π de ordem ξ. Para tal irá ser utilizado o seguinte teorema de Barbour [1].

Lema 1: Seja uma cadeia de Markov S com distribuição estacionária μ e matriz de probabilidades de transição Π e seja V o número de visitas ao conjunto T até ao instante n. Então com $\lambda = E(V)$ tem-se:

$$d(V, Po(\lambda)) \leq (1 - e^{-\lambda}) \left[\mu(T) + \frac{2}{\mu(T)} \sum_{r,s \in T} \mu_r \sum_{j \geq 1} \left| \Pi_{r,s}^{(j)} - \mu_s \right| \right] \tag{4.5}$$

4.2.1 Cadeia de Markov de dois estados

Teorema 1: Consideremos uma cadeia de Markov estacionária $\{S_j\}_{1 \leq j \leq n}$ tomando valores no alfabeto $\mathcal{A} = \{I, S\}$, com matriz de probabilidades de transição definida por:

$$\Pi = \begin{array}{c} \\ \text{S} \\ \text{I} \end{array} \begin{array}{c} \begin{array}{cc} \text{S} & \text{I} \end{array} \\ \begin{bmatrix} \alpha & 1\text{-}\alpha \\ \beta & 1\text{-}\beta \end{bmatrix} \end{array} \quad (0 < \beta < \alpha < 1)$$

e probabilidades estacionárias dadas por $p = P(S) = \frac{\beta}{1-\alpha+\beta}$ e $q =$

$P(I) = 1 - p = \frac{1-\alpha}{1-\alpha+\beta}$

Sendo N, A e B definidos como no caso i.i.d., então:

$$d(N, Po(E(A)) \leq \frac{n\beta\alpha^{2k-1}}{1-\alpha+\beta} + \beta\alpha^{k-1}\left(\frac{1+(2k+1)(1-\alpha)}{1-\alpha+\beta} + \frac{2(\alpha-\beta)(1-\alpha)}{(1-\alpha+\beta)^2}\right) \tag{4.6}$$

Demonstração: Utilizando as variáveis A e B tem-se, como no caso i.i.d:

$$d(B, N) \leq P(B \neq N) \leq \frac{n\beta\alpha^{2k-1}}{1-\alpha+\beta} \tag{4.7}$$

$$d(A, B) \leq P(A \neq B) \leq \frac{\beta\alpha^{k-1}}{1-\alpha+\beta} \tag{4.8}$$

Utilizando o **lema 1** e sendo $T = \underset{|---k\,S's----|}{\{ISSS...SSS\}} = \{r\}$ vem

$$d(A, Po(E(A)) \leq \left[\frac{(1-\alpha)\beta\alpha^{k-1}}{1-\alpha+\beta} + 2\sum_{j\geq 1}\left|\Pi_{r,r}^{(j)} - \frac{(1-\alpha)\beta\alpha^{k-1}}{1-\alpha+\beta}\right|\right] \tag{4.9}$$

Note-se que $\Pi_{r,r}^{(j)} = 0$ se $j \leq k$ e para $j \geq k+1$,

$$\Pi_{r,r}^{(j)} = \Pi_{S,I}^{(j-k)}\beta\alpha^{k-1} \tag{4.10}$$

Como $\Pi^{(i)} = \begin{bmatrix} p+q(\alpha-\beta)^i & q(1-(\alpha-\beta)^i) \\ p(1-(\alpha-\beta)^i) & q+p(\alpha-\beta)^i \end{bmatrix}$ de (4.9) e (4.10) e dado $\alpha > \beta$,

$$\begin{aligned} d(A, Po(E(A)) &\leq \frac{(2k+1)(1-\alpha)\beta\alpha^{k-1}}{1-\alpha+\beta} + 2\sum_{j\geq k+1}\frac{(1-\alpha)\beta\alpha^{k-1}(\alpha-\beta)^{j-k}}{1-\alpha+\beta} \\ &= \frac{(2k+1)(1-\alpha)\beta\alpha^{k-1}}{1-\alpha+\beta} + \frac{2(\alpha-\beta)(1-\alpha)\beta\alpha^{k-1}}{(1-\alpha+\beta)^2} \end{aligned} \tag{4.11}$$

O resultado segue por (4.7), (4.8) e (4.11). Para $\alpha < \beta$ o resultado seria similar.

No resultado anterior a aproximação é de ordem $O(\max\{k\beta\alpha^{k-1}, \alpha^k\})$, com a condição de que $\frac{n\beta\alpha^{k-1}}{1-\alpha+\beta} \underset{n\to\infty}{\longrightarrow} \lambda$

Geske[5] provou, por um processo análogo, que para α grande e β pequeno tem-se que:

$$d(M, Po(E(M)) \leq \frac{2\alpha}{1-\alpha} + \frac{\beta\alpha^{k-1}}{1-\alpha+\beta} + 2\frac{(1-\alpha)(\alpha-\beta)\alpha^{k-1}}{(1-\alpha+\beta)^2}$$

sendo $M(n,k)$ o número de cadeias de k sucessos contados com sobreposição.

4.2.2 Cadeia de Markov generalizada

Se considerarmos agora uma palavra $W = (w_1...w_k)$ de comprimento k e quisermos aproximar N por uma variável de Poisson, então a situação torna-se um pouco mais complicada e precisaremos de definir novamente duas variáveis A e B. Para tal iremos ainda necessitar de falar em extensões da palavra W.

Relembramos ainda, que uma palavra $W = (w_1...w_k)$ diz-se m-repetitiva se as primeiras m letras são iguais e na mesma ordem às últimas m. Uma palavra pode ser $m_1 < m_2 < ... < m_r = m$ repetitiva, com $1 \leq m \leq k-1$.

Extensões

Uma extensão da palavra W é uma sequência da forma $C_1C_2...C_qWD_1...D_p$ onde para cada C_i $(1 \leq i \leq q)$ de $k - m_j$ $(1 \leq j \leq r)$ letras, o grupo de k letras começando com o primeiro caracter de C_i é uma ocorrência da palavra e onde para cada grupo D_i $(1 \leq i \leq p)$ de $k - m_j$ letras, o grupo de k letras que acabam com o último D_i é uma ocorrência da palavra.

É evidente que a representação de uma extensão em termos de $C's$ e $D's$ não é única.

Exemplos:
ABRACAD**ABRACADABRA**BRACADABRA
(com C_1 =ABRACAD e D_1 =BRACADABRA)
e **ABRACADABRA**CADABRABRACADABRA
(com D_1 =CADABRA e D_2 =BRACADABRA)
são extensões da palavra
$W = ABRACADABRA$ que é uma palavra m_1 e m_4 repetitiva.
Vamo-nos referir a esta última, do tipo $WD_1...D_i$ $(1 \leq i \leq p)$ como sendo a forma canónica de uma extensão da palavra W.

Uma extensão diz-se maximal e que termina na posição j num grupo de n letras se:
1. uma extensão canónica da forma $WD_1...D_p$ termina na posição j.
2. para nenhum i $(1 \leq i \leq r)$ a palavra W começando com $k-m_j$ letras antes do início da extensão canónica coincide com W.
3. A palavra W não ocorre na m_i-ésima posição qualquer que seja i $(1 \leq i \leq r)$
É evidente que no caso de cadeias de sucessos uma extensão maximal é meramente a ocorrência de uma cadeia de sucessos de comprimento pelo menos k.

Seja B o número de extensões maximais da palavra W numa sequência de n letras.
De seguida defina-se:
A_1- n^{o}de ocorrências das palavras da forma $w_2w_3...w_kw_1...w_k$
A_2- n^{o}de ocorrências das palavras da forma $w_3...w_k\_w_1...w_k$
$\vdots$
A_i- n^{o}de ocorrências das palavras da forma $w_{i+1}w_{i+2}...w_k\_...\_w_1...w_k$
$\vdots$
$A_{k-1}-$ n^{o}de ocorrências das palavras da forma $w_k\_...\_w_1...w_k$
A_k- n^{o}de ocorrências das palavras da forma $\_...\_w_1...w_k$
onde $\_$ representa uma posição em branco (em cada categoria i existem $i-1$ posições em branco) que pode ser preenchida por qualquer letra do alfabeto desde que siga as regras:

1. para cada $j(1 \leq j \leq r)$ as $k - m_j$ letras que precedem as últimas k $(w_1...w_k)$ não conduzem a uma extensão C_i à esquerda.

2. para cada i, apenas as palavras que não aparecem em $A_j(j < i)$ são consideradas em A_i.

Finalmente seja $A = \sum_{i=1}^{k} A_i$ o número de ocorrências, contadas com sobreposição, de $2k-1$ letras que terminam em $W = (w_1...w_k)$ e não a extendem à esquerda, i.e, para as $2k-1$ letras mencionadas anteriormente apenas existe uma ocorrência de W.

Por simplicidade usaremos a notação A_i para nos referirmos às palavras na categoria i assim como ao número de ocorrências repetitivas de tais palavras.

No caso de cadeias de sucessos A é meramente o número de ocorrências de $IS...S$ $(k\ S's)$

Exemplo: Seja $W = ALFA$ e $\mathcal{A} = \{A, F, L\}$ então as quatro categorais definidas anteriormente são:

$A_1 = LFAALFA$ (1 palavra possível)

$A_2 = FA\_ALFA$ (3 palavras possíveis)

$A_3 = A\_\ \_ALFA$ (8 palavras possíveis, não pode ser $ALFALFA$)

$A_4 = \_\ \_\ \_\ ALFA$ (27-1-3-8-1 palavras possíveis, excluem-se as anteriores assim como $ALFALFA$)

Tem-se então os seguintes resultados:

1. As palavras A_i são $k - i$ repetitivas.

2. Uma palavra A_j pode sobrepor uma palavra A_i (da esquerda para a direita da última) no máximo em $k - i$ letras

3. Uma palavra A_j pode sobrepor uma palavra A_i (da direita para a esquerda da última) no máximo em $k - j$ letras

Lema 2: $A = B$ se a palavra W não está correctamente "escrita" em alguma das $k - 1$ posições.

Demonstração: Assumindo que a palavra não está correctamente escrita em alguma das $k - 1$ posições. É evidente que $A = 0$ sse $B = 0$.

Assumindo que $B \geq 1$, consideremos uma extensão maximal a começar na posição j ($j \geq k$). É evidente que uma palavra A é obtida terminando na posição $j + k - 1$ (se uma palavra A termina na posição $j + k - 1$ uma extensão maximal deve ter começado na posição j) e também que uma palavra A não poder ser obtida em qualquer outra posição dentro da extensão maximal. Finalmente uma palavra A que termina depois do fim (ou antes do início) da extensão maximal em questão deve corresponder a uma outra extensão maximal.

Lema 3: A probabilidade de uma extensão maximal conter duas ou mais ocorrências repetitivas da palavra W é menor que kp^{2k}

Demonstração: Uma extensão maximal deve começar com a palavra W.

Consideremos que a primeira ocorrência não repetitiva de W depois da inicial e denominemos por δ ($0 \leq \delta \leq k - 1$) o número de espaços entre o fim da primeira ocorrência e o início da segunda. A probabilidade dessa extensão maximal é no máximo p^{2k}.

Suponhamos que $\delta \geq k$, então a ocorrência deve terminar com um segmento do tipo D de comprimento $k - m_j$. As k letras que precedem este segmento constituem uma ocorrência anterior de W que é disjunta da anterior o que é uma contradição.

Iremos então agora aproximar a variável aleatória N por uma variável com distribuição de Poisson, mas para tal iremos necessitar de alguns resultados sobre a matriz de probabilidades de transição Π. Esta não é geralmente reversível mas, Fiil(1991) [3], definiu a sua reversibilidade multiplicativa $M(\Pi)$ por $M(\Pi) = \Pi\Pi^*$, onde Π^* é definida por:

$$\Pi^*_{x,y} = \frac{\mu_y \Pi_{y,x}}{\mu_x}.$$

Fill [3] mostrou ainda que os valores próprios de $M(\Pi)$ são todos reais e não negativos e denominando $\delta = \delta(M)$ como o segundo maior valor próprio de $M(\Pi)$,

$$\sup_y |\Gamma_{x,y}^{(n)} - \mu_y| \leq \frac{1}{2}\frac{\delta^{n/2}}{\sqrt{\pi_x}} \tag{4.12}$$

Teorema 2: Seja N o número de ocorrências não repetitivas de uma palavra fixa $W = (w_1 w_2 ... w_k)$ de comprimento k, obtida recorrendo uma cadeia de Markov $\{W_j\}_{1 \leq j \leq n}$, com alfabeto $\mathcal{A}$ com ξ letras, matriz de probabilidades de transição Π, probabilidade estacionaria μ e o segundo maior valor próprio da reversibilidade multiplicativa de Π denominado por δ. Seja A o número de ocorrências repetitivas de palavras de comprimento $(2k-1)$ que terminam em $w_1 w_2 ... w_k$ mas que contêm apenas uma ocorrência da palavra. Então:

$$\begin{aligned} d(N, Po(E(A)) \quad \leq \quad & k\boldsymbol{\mu}(W)\left\{1 + \frac{2}{1 - \max\limits_{x,y}\Pi_{x,y}} + n\boldsymbol{\mu}(W)\right\} \\ & + \boldsymbol{\mu}(W)\frac{\sqrt{\delta}}{(1-\sqrt{\delta}}\frac{1}{\sqrt{\mu_{w_k}}\mu_{w_1}} \\ & \times \left\{k + \xi - 1 + \frac{1}{1 - \max\limits_{x,y}\Pi_{x,y}} + \frac{n\boldsymbol{\mu}(W)}{2}\right\} \end{aligned} \tag{4.13}$$

Demonstração: Como anteriormente, consideremos B o número de extensões maximais da palavra W e A_{ij} qualquer palavra na categoria A_i, Então pelo **Lema 2**:

$$d(A, B) \leq P(A \neq B) \leq (k-1)\boldsymbol{\mu}(W) \tag{4.14}$$

e como $B \neq N$ se uma extensão maximal de W contém duas ou mais ocorrências não repetitivas da palavra W, pelo **Lema 3**,

$$d(N, B) \leq P(N \neq B) \leq n\boldsymbol{\mu}(W)\sum_{r=1}^{k}\Pi_{w_k,w_1}^{(r)}\Pi_{w_1,w_2}...\Pi_{w_{k-1},w_k} \tag{4.15}$$

Utilizando (4.12) tem-se

$$\begin{aligned} d(N,B) &\leq n\boldsymbol{\mu}(W)^2 \sum_{r=1}^{k} \frac{\Pi^{(r)}_{w_k,w_1} - \mu_{w_1} + \mu_{w_1}}{\mu_{w_1}} \\ &\leq n\boldsymbol{\mu}(W)^2 \left\{ k + \sum_{r=1}^{k} \frac{\left|\Pi^{(r)}_{w_k,w_1} - \mu_{w_1}\right|}{\mu_{w_1}} \right\} \\ &\leq n\boldsymbol{\mu}(W)^2 \left\{ k + \frac{1}{2} \sum_{r=1}^{k} \frac{\delta^{r/2}}{\mu_{w_1}\sqrt{\mu_{w_k}}} \right\} \\ &\leq n\boldsymbol{\mu}(W)^2 \left\{ k + \frac{1}{2} \frac{\sqrt{\delta}}{(1-\sqrt{\delta})\mu_{w_1}\sqrt{\mu_{w_k}}} \right\} \end{aligned} \tag{4.16}$$

Retomemos agora a comparação entre A e $Po(E(A))$ e denotemos $\boldsymbol{\mu}(A)$ e $\boldsymbol{\mu}_{i,j}$ como a probabilidade estacionária de uma palavra A e de uma palavra específica A_{ij} respectivamente. Utilizando o **Lema 1** tem-se:

$$d(A, Po(E(A)) \leq (1-e^{-E(A)}) \left(\boldsymbol{\mu}(A) + \frac{2}{\boldsymbol{\mu}(A)} \sum_{i,j} \boldsymbol{\mu}_{i,j} \sum_{l,m} \sum_{r\geq 1} \left| \Pi^{(r)}_{ij,lm} - \boldsymbol{\mu}_{l,m} \right| \right) \tag{4.17}$$

Vamos separar a soma em duas partes, uma para $r < 2k-1$ e outra para $r \geq 2k-1$. Relembremos ainda que A_{lm} pode sobrepor A_{ij} (da direita para a esquerda da última) em pelo menos $k-l$ palavras.

A soma para $r < 2k-1$ na verdade apenas extende em *r's* no intervalo $\{k+l-1, ..., 2k-2\}$.

Denominando as palavras A_{ij} e A_{lm} por

$w_{i+1,}w_{i+2}...w_k d_1...d_{i-1}w_1w_2...w_k$ e $w_{l+1,}w_{l+2}...w_k c_1...c_{l-1}w_1w_2...w_k$ respectivamente, observa-se que para cada $r \in \{k+l-1, ..., 2k-2\}$ e $l \leq k-1$ (para $l = k$ a soma é zero)

$$\left| \Pi^{(r)}_{ij,lm} - \boldsymbol{\mu}_{l,m} \right| = \Pi^{(r)}_{ij,lm} - \boldsymbol{\mu}_{l,m} \leq \Pi^{(r)}_{ij,lm} \tag{4.18}$$

e então:

$$\sum_{r=k+l-1}^{2k-2} \left| \Pi_{ij,lm}^{(r)} \; - \; \boldsymbol{\mu}_{l,m} \right| \leq$$

$$\begin{aligned}
\leq \; & \Pi_{w_k,w_{l+2}} \Pi_{w_{l+2},w_{l+3}} ... \; \Pi_{w_k,c_1} \Pi_{c_1,c_2} ... \Pi_{c_{l-2},c_{l-1}} \Pi_{c_{l-1},w_1} \Pi_{w_1,w_2} ... \Pi_{w_{k-1},w_k} \\
& + \Pi_{w_k,w_{l+3}} \Pi_{w_{l+3},w_{l+4}} ... \; \Pi_{w_k,c_1} \Pi_{c_1,c_2} ... \Pi_{c_{l-2},c_{l-1}} \Pi_{c_{l-1},w_1} \Pi_{w_1,w_2} ... \Pi_{w_{k-1},w_k} \\
& + ... + \Pi_{w_k,c_1} \Pi_{c_1,c_2} ... \Pi_{c_{l-2},c_{l-1}} \Pi_{c_{l-1},w_1} \Pi_{w_1,w_2} ... \Pi_{w_{k-1},w_k} \\
\leq \; & \Pi_{w_k,c_1} \Pi_{c_1,c_2} .. \Pi_{c_{l-2},c_{l-1}} \Pi_{c_{l-1},b_1} \Pi_{b_1,b_2} ... \Pi_{b_{k-1},b_k} \times \\
& \times \left\{ 1 + \max_{x,y} \Pi_{x,y} + \left(\max_{x,y} \Pi_{x,y} \right)^2 + ... \right\} \\
\leq \; & \frac{\Pi_{w_k,c_1} \Pi_{c_1,c_2} ... \Pi_{c_{l-2},c_{l-1}} \Pi_{c_{l-1},w_1} \Pi_{w_1,w_2} ... \Pi_{w_{k-1},w_k}}{1 - \max_{x,y} \Pi_{x,y}}
\end{aligned} \tag{4.19}$$

Somando em m e l, tem-se

$$\begin{aligned}
\sum_{l,m} \sum_{r=k+l-1}^{2k-2} \left| \Pi_{ij,lm}^{(r)} \; - \; \boldsymbol{\mu}_{l,m} \right| \; & \leq \; \sum_{1 \leq l \leq k} \frac{\Pi_{w_k,w_1}^{(l)} \Pi_{w_1 w_2} ... \Pi_{w_{k-1},w_k}}{1 - \max_{x,y} \Pi_{x,y}} \\
& \leq \; \boldsymbol{\mu}(W) \; \frac{\sum_{l=1}^{k} \Pi_{w_k,w_1}^{(l)}}{\mu_{w_1}} \frac{1}{1 - \max_{x,y} \Pi_{x,y}} \\
& \leq \; \boldsymbol{\mu}(W) \left\{ k + \frac{1}{2} \frac{\sqrt{\delta}}{(1 - \sqrt{\delta})} \frac{1}{\sqrt{\mu_{w_k}} \mu_{w_1}} \right\} \times \\
& \quad \times \frac{1}{1 - \max_{x,y} \Pi_{x,y}}
\end{aligned} \tag{4.20}$$

A contribuição desta parcela para (4.17) é no máximo de

$$2\mu(W) \left\{ k + \frac{1}{2} \frac{\sqrt{\delta}}{(1 - \sqrt{\delta})} \frac{1}{\sqrt{\mu_{w_k}} \mu_{w_1}} \right\} \frac{1}{1 - \max_{x,y} \Pi_{x,y}} \tag{4.21}$$

Consideremos agora o caso em que $r \geq 2k-1$ e observando que para esses valores de r e para $l \leq k-1$,

$$\sum_{r\geq 2k-1} \left|\Pi^{(r)}_{ij,lm} \; - \; \boldsymbol{\mu}_{l,m}\right| \leq$$

$$\begin{aligned}
\leq \;& \Pi_{w_{l+1},w_{l+2}}...\Pi_{w_{k-1},w_k}\Pi_{w_k,c_1}\Pi_{c_1,c_2}...\Pi_{c_{l-2},c_{l-1}}\Pi_{c_{l-1},w_1}...\Pi_{w_{k-1},w_k} \times \\
& \times \sum_{r\geq 2k-1} \left|\Pi^{(r-2k+2)}_{w_k,w_{l+1}} \; - \; \mu_{w_{l+1}}\right| \\
\leq \;& \Pi_{w_{l+1},w_{l+2}}...\Pi_{w_{k-1},w_k}\Pi_{w_k,c_1}\Pi_{c_1,c_2}...\Pi_{c_{l-2},c_{l-1}}\Pi_{c_{l-1},w_1}...\Pi_{w_{k-1},w_k} \times \\
& \times \sum_{r\geq 2k-1} \frac{\left(\sqrt{\delta}\right)^{r-2k+2}}{2\sqrt{\mu_{w_k}}} \\
= \;& \frac{1}{2}\frac{\sqrt{\delta}}{(1-\sqrt{\delta})}\frac{1}{\sqrt{\mu_{w_k}}}\Pi_{w_{l+1},w_{l+2}}...\Pi_{w_{k-1},w_k}\Pi_{w_k,c_1}\Pi_{c_1,c_2}...\Pi_{c_{l-1},w_1}...\Pi_{w_{k-1},w_k}
\end{aligned}$$

Somando em m, obtém-se para $l \leq k-1$

$$\sum_m \sum_{r\geq 2k-1} \Pi^{(r)}_{ij,lm} \; - \; \boldsymbol{\mu}_{l,m} \leq \frac{1}{2}\frac{\sqrt{\delta}}{(1-\sqrt{\delta})}\frac{1}{\sqrt{\mu_{w_k}}}\frac{\Pi_{w_{l+1},w_{l+2}}\Pi_{w_{k-1},w_k}}{\mu_{w_1}}\Pi^{(l)}_{w_k,w_1}\boldsymbol{\mu}(W) \tag{4.22}$$

Por outro lado para $l = k$,

$$\begin{aligned}
\sum_{r\geq 2k-1} \left|\Pi^{(r)}_{ij,km} \; - \; \boldsymbol{\mu}_{k,m}\right| \;\leq\; & \Pi_{c_1,c_2}...\Pi_{c_{k-2},c_{k-1}}\Pi_{c_{k-1},w_1}\Pi_{w_1,w_2}...\Pi_{w_{k-1},w_k} \times \\
& \times \sum_{r\geq 2k-1} \left|\Pi^{(r-2k+2)}_{w_k,c_1} \; - \; \mu_{c_1}\right| \\
= \;& \frac{1}{2}\frac{\sqrt{\delta}}{(1-\sqrt{\delta})}\frac{1}{\sqrt{\mu_{b_k}}}\frac{\Pi_{c_1,c_2}...\Pi_{c_{k-2},c_{k-1}}\Pi_{c_{k-1},w_1}\mu(W)}{\mu_{w_1}}
\end{aligned}$$

que somando em m, dá:

$$\sum_{m} \sum_{r \geq 2k-1} \left| \Pi_{ij,km}^{(r)} \; - \; \boldsymbol{\mu}_{k,m} \right| \leq \frac{\xi \boldsymbol{\mu}(W)}{\mu_{w_1}} \frac{\sqrt{\delta}}{2(1-\sqrt{\delta})} \frac{1}{\sqrt{\mu_{w_k}}} \tag{4.23}$$

uma vez que $\sum_{m=1}^{\xi} \Pi_{c_1,c_2} ... \Pi_{c_{k-2},c_{k-1}} \Pi_{c_{k-1},w_1} \leq \sum_{m=1}^{\xi} 1 \leq \xi$

As expressões (4.22) e (4.23) mostram que a contribuição para (4.17) dos termos $r \geq 2k-1$ é majorado por:

$$\boldsymbol{\mu}(W) \frac{\sqrt{\delta}}{(1-\sqrt{\delta})} \frac{1}{\sqrt{\mu_{w_k}} \mu_{w_1}} \{k - 1 + \xi\} \tag{4.24}$$

Utilizando (4.20) e (4.23) temos então que:

$$\begin{aligned} d(A, Po(E(A)) \quad \leq \quad & \boldsymbol{\mu}(A) + \frac{2k\boldsymbol{\mu}(W)}{1 - \max\limits_{x,y} \Pi_{x,y}} + \boldsymbol{\mu}(W) \frac{\sqrt{\delta}}{(1-\sqrt{\delta})} \frac{1}{\sqrt{\mu_{w_k}} \mu_{w_1}} \times \\ & \times \left\{ k + \xi - 1 + \frac{1}{1 - \max\limits_{x,y} \Pi_{x,y}} \right\} \end{aligned} \tag{4.25}$$

O resultado segue por (4.14), (4.16) e (4.25).

Estes resultados podem ser utilizados para encontrar majorantes do erro cometido quando se utiliza uma aproximação à distribuição de Poisson. Actualmente utilizam-se muitos destes resultados em áreas como a criptografia e a análise de sequências de ADN.

Capítulo 5

Conclusão

Neste trabalho procuramos reunir, tanto quanto possível, resultados importantes no estudo das ocorrências de palavras em sequências aleatórias de letras. A principal contribuição deste trabalho foi a aplicação dos resultados ao estudo de sequências do ADN. Outras aplicações poderiam ter sido pensadas mas, não é fascinante descobrir o que nos torna distintos uns dos outros? Que relações existem entre sequências do ADN? A todas estas questões os biólogos tentam responder mas, para isso necessitam que os matemáticos desenvolvam novos modelos, novas abordagens, para melhor modelar as sequências de ADN.

Assim pensamos que os objectivos iniciais a que nos propusemos foram cumpridos, no entanto existe a consciência de que apenas foi estudado uma ínfima parte do problema.

Assim num trabalho futuro, poderíamos estudar sequências de ADN especiais no que concerne à sua função biológica e que tivessem um comportamento inesperado. Para alguns autores estas sequências são designadas por "motivos estruturais". Poderíamos estudar a probabilidade de ocorrência de tais motivos e também aproximá-los por uma distribuição de Poisson composta, tal como tem sido desenvolvido por alguns autores, entre os quais Robin [8].

Capítulo 6

Apêndice

Para determinar a variância é necessário fazer a decomposição da matriz Π como anteriormente referimos. Assim utilizando o programa *Mathematica* para fazer essa decomposição (relembramos que a Matriz A será a matriz dos vectores próprios (em coluna); Λ a matriz diagonal, cujos elementos serão os valores próprios e B a inversa de A) tem-se, com aproximação de quatro dígitos significativos:

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & .06847 + .06525i & 0 & 0 \\ 0 & 0 & .06847 - .06525i & 0 \\ 0 & 0 & 0 & -.09448 \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & -.3825 - .8191i & -.3825 + .8191i & -.6862 \\ 1 & -.1956 + .43929i & -.1956 - .43929i & -2.934 \\ 1 & -.3847 + .2834i & -.3847 - .2834i & 2.877 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

$$B = \begin{pmatrix} .2341 & .274 & .2518 & .24 \\ -.1028 + .4076i & -.0562 - .261i & -.21 - .181i & .3689 + .035i \\ -.1028 - .4076i & -.0562 + .261i & -.21 + .181i & .3689 - .035i \\ -.02855 & -.1617 & .1681 & .02212 \end{pmatrix}$$

Desta forma

$$\begin{aligned}\gamma_{k,1}^{'}(1) &= \frac{-1}{.274}\left[\begin{array}{c}\frac{.068+.065i}{1-.0068-.065i}\times(-.385+.283i)\times(-0.056-.26i)+\\ +\frac{.068-.065i}{1-.068+.065i}\times(-.385-.28)\times(-.056+.26i)+\\ \frac{-.0984}{1+.0984}\times.2887\times(-.16)\end{array}\right]\\ &= 0.0219991+0.0061888i\end{aligned}$$

Bibliografia

[1] Barbour,A. D., Host, L. e Janson, S. (1992). *Poisson Aproximation*. Oxford University Press.

[2] Feller, W. (1968). *An Introdution to Probability and its Applications*, Vol I, 3rd edn. New York, John Wiley

[3] Fill, J. A. (1991). Eigenvalue bounds on convergence to stationarity for nonreversible Markov Chains, with an application to the exclusion process. *Ann. Appl. Prob.* **1**, 62-87

[4] Fu, J. C. e Koutras, M. V. (1994). Distribution theory of runs: a Markov Chain approch, J. *Amer. Statist. Assoc.*, **89**, 1050-1058

[5] Geske, M. X., Godbole, A. P., Schaffner, A. A., Skolnick, A. M. e Wallstrom, G. L. (1995). Compound Poisson approximation for word patterns under Markovian hypotesis. *J. Appl. Prob.* **32**, 877-893

[6] Godbole, A. e Schaffner, A. (1993). Improved Poisson aproximation for word patterns. *Adv. Appl. Prob.* **25**, 334-347

[7] Karlin, S. (1975). *A First Course in Stochastic Processes*, Academic Press

[8] Robin, S. e Daudin, J. J. (1999). Exact distributions of word ocurrences in a random sequence of letters. *J. Appl. Prob.* **36**, 179-193